A METHOD AND SYSTEM FOR NORMALIZING DIRTY TEXT IN A DOCUMENT

ABSTRACT OF THE INVENTION

A method and system of normalizing dirty text in a document. The present invention creates a thesaurus that evolves over time as new document collections are analyzed. This thesaurus, which is used by an editor, contains standard terms and phrases, and their corresponding variations of these standard terms and phrases. Documents are run through this editor and misspelled words or phrases, joined words, and ad hoc abbreviations are replaced with standard terms from the thesaurus. The present invention also enables normalization of documents in cases where a list of standard terms must be inferred from the corpus of the document. The normalizer will facilitate data mining applications which can not function properly with dirty text, resulting in more accurate analysis of documents. Over time, as the thesaurus evolves, collecting more words and phrases, the process of generating the thesaurus will become more automated.